The “cocktail-party effect” and prosodic rhythm: Discrimination of the temporal structure of speech-like sequences in temporal interference

Pierre L. Divenyi[1] **and Alex Brandmeyer**[2]
Veterans Affairs Northern California Health Care System and East Bay Institute for Research and Education
Martinez, CA 94553, U.S.A.
E-mail: pdivenyi@ebire.org[1], abrandmeyer@ebire.org[2]

ABSTRACT

A dimension crucial for the perceptual segregation of simultaneous speech sounds is the one of syllabic and subsyllabic envelope fluctuations, indicating that, in the “cocktail-party effect,” prosody plays a definite role. This role was investigated in a psychoacoustic study using a much reduced stimulus configuration to examine the way the rhythmic pattern of a target stream can be segregated from another, irrelevant stream. Both streams consisted of 40-ms bursts of harmonic complexes with different fundamental frequencies and the same average modulation rate. The target stream had a three-burst dactyl or amphibrach rhythm, whereas the interfering stream’s burst rhythmic pattern was random. Results showed that a higher-f_0 target mixed with a lower-f_0 interference was significantly easier to identify than the opposite, and that difficulty of the task increased with modulation rate. Elderly subjects performed significantly poorer than the young, for both segregating streams based on prosodic rhythm and understanding speech in babble.

1. INTRODUCTION

Prosody is much more than once thought: It does not merely represent a dimension that confers supplementary nonlinguistic information to utterances but, through an intricate set of language-specific rules organically connected to the phonemic rules, it actively participates in forming meaning [1]. It is, therefore, unsurprising that speech perception is much influenced by prosody: incorrect or inappropriate stress accent, syllable duration, or pitch contour will harm intelligibility [2]. Among prosodic cues, syllable duration, i.e., prosodic rhythm, acquires a special role because it signals not only semantic categories [3] but also syntactic boundaries [4]. When speech is presented in an unfavorable signal-to-noise ratio, especially when the noise consists of speech babble, the fluctuations of the speech envelope become attenuated inmost frequency bands [5]. In such “cocktail-party” situations, syllabic and sub-syllabic duration cues are among the features of the target speech that will become masked. Previous work in our laboratory [6] identified syllabic-rate rhythm as one of the cardinal dimensions of the process of perceptual segregation of simultaneous sources, or “streams” [7]. Segregation of streams with different prosodic-like rhythmic patterns has been also shown to become difficult for persons reduced ability to understand speech in a noisy background, such as the elderly [8]. The present paper reports on results of a series of experiments that investigated perceptual segregation of two simultaneous streams of sounds which, although non-speech, through the choice of parameters, emulated a “cocktail-party” situation of a relevant target stream and an irrelevant interference, or distracter, stream.

2. METHODS

The objective of the study was to determine the precise ability to discriminate syllabic and subsyllabic rhythmic patterns in one of the streams designated as *target* in the presence of another stream in which the rhythmic information was random. Specifically, the target stream consisted of a complex tone having fundamental frequency $f_{0\mathrm{Targ}}$ Hz and equal-amplitude harmonic components between 600 and 3000 Hz; this complex tone was amplitude-modulated with a waveform to produce three bursts of 40 ms total duration that included a 10-ms rise and a 10-ms fall time. The interval between the onsets of the first and the last burst was $\boldsymbol{T}$ ms. The two intervals separating the onsets of the three consecutive bursts had an average duration $\boldsymbol{T}/2$, with one of the intervals being longer by τ ms and the other shorter by τ ms. In psychoacoustic terms, the average modulation frequency of the target was $2/\boldsymbol{T}$ Hz. In terms of poetry feet, the long-short inter-burst duration pattern was a dactyl and the short-long pattern an amphibrach. At each trial of a 2-AFC (two-alternative forced choice) block, the listener heard both patterns but in a random order; he/she had to decide by pressing one of two buttons what the order was. The duration of the interval τ was modified within each block of trials according to a modified up-down procedure [9] to track the value of τ at which the rhythmic patterns in the target stream were discriminable at the 71 % correct level, which we considered to be the rhythmic discrimination threshold.

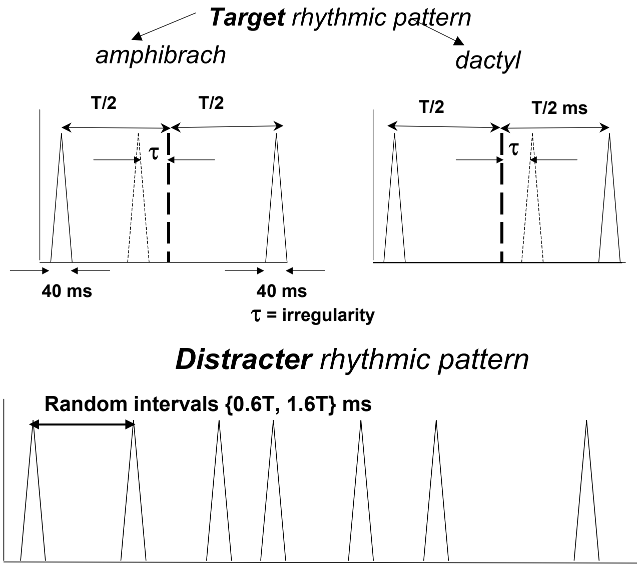

Figure 1. Schematic time diagram of the stimulus. *Top half of figure*: diagram of the target pattern consisting of three identical bursts (including the 10-ms cosine-square on- and off-ramps). At each trial, either the amphibrach rhythmic pattern on the left, or the dactyl rhythmic pattern, on the right, was presented. The pattern was generated by making the temporal position of the middle burst appear either advanced or delayed by τ ms with respect to a rhythmically regular tribrach pattern (illustrated by the heavy dashed line). The interval separating the onsets of the first and the last bursts was ***T*** ms. *Bottom half of figure*: diagram of the distracter rhythmic pattern consisting of bursts of sinusoidal harmonic complexes also having a nominal duration of 40-ms but a different fundamental frequency f_{0Dist}. The intervals separating consecutive bursts in the distracter stream were random variables with a rectangular distribution chosen from the range of {0.6 ***T***, 1.6***T***} ms, i.e., 40% shorter to 40% longer than the 200 ms interval at which the target stream would be rhythmically regular. The onset and the offset of the distracter stream was also random with respect to that of the target stream with the constraint that random sampling should start 250 ms before the first and end 250 ms after the last target burst.

Simultaneously with the target stream, an interfering *distracter* stream consisting of a different amplitude-modulated complex tone having fundamental frequency f_{0Dist} was also presented. The frequency range of this tone was also between 600 and 3000 Hz and its average modulation frequency 2/***T***. The distracter was amplitude-modulated by a signal that produced a series of 40-ms bursts with random inter-burst intervals rectangularly distributed in log time between 0.6***T*** and 1.6***T***. The distracter stream began 300 ms before and ended 300 ms after the target stream. The target and the distracter are diagrammatically illustrated in Figure 1.

Several stimulus conditions were investigated. First, two different degrees of separation between the target and distracter fundamental frequencies were studied (27 and 77 percent, corresponding to the approximate musical intervals of Major third and minor seventh, respectively), with the lower of the two frequencies fixed at 107 Hz. Second, assignment of the target and distracter patterns to the two streams was also a parameter: in one configuration the target had the higher and the distracter the lower, 107-Hz fundamental, whereas in the other configuration the target fundamental frequency was 107 Hz and the distracter the higher. Third, we investigated three average modulation frequencies $f_{ModAver}$: 5, 10, and 20 Hz. Finally, rhythmic discrimination was studied at two modulation depths: 100% (as shown in Figure 1) and 50% (i.e., with a pedestal 6 dB below the burst peaks). Throughout the study, the target and the distracter streams were presented at identical spectral levels of 70 dB SPL at the subject's earphones. All stimuli were presented to both ears at identical levels.

In a control condition, the target stream alone was presented – this condition determined the limits of rhythmic discrimination. In contrast, we refer to the other conditions, in which the target and the distracter were simultaneously presented, as those of stream segregation because the discrimination task could only be performed if the listener correctly identified the target stream and perceptually segregated it from the distracter stream. Prior to each block of segregation trials, the listener was cued to the target stream by presenting an unmodulated burst of complex tone having the appropriate f_0. five young (18 to 30 years of age) and seven elderly listeners (60 to 76 years of age) participated.

3. RESULTS

Experiment 1 investigated the effect of fundamental frequency of the target pattern and the distracter stream on the discriminability of the pattern rhythm, i.e., the minimum deviation from a perfectly regular train of three pulses at which the subject could identify the pattern as either an amphibrach (short-long) or a dactyl (long-short). Data for the 400-ms total pattern duration, i.e., for the average amplitude modulation rate of 5 Hz, are shown in Figure 2 for five young and seven elderly listeners, for the condition in which the pattern stream f_0 was higher than that of the distracter stream, and for the condition in which it was lower. For comparison, data for the discrimination of the pattern alone, i.e., without the distracter, are also shown. This experiment is analogous to examining the effect of the gender of a target taker and a distracter talker on the perception of prosody, and the effect of the target talker having a higher- or lower-pitched voice than the distracter talker.

In Experiment 2, the effect of changing the modulation depth from 100 percent (i.e., the one shown in Figure 1) to 50 percent on the discriminability of the pattern rhythm was investigated. Data for the 400-ms total pattern duration, the largest fundamental frequency difference (76 %), and the condition with the target stream f_0 always higher than the distracter stream f_0 are shown in Figure 3 for the young and the elderly subjects. This experiment is analogous to examining the effect of speaking style on the perception of prosody, with two simultaneous talkers speaking either in a well-articulated or in a colloquial style.

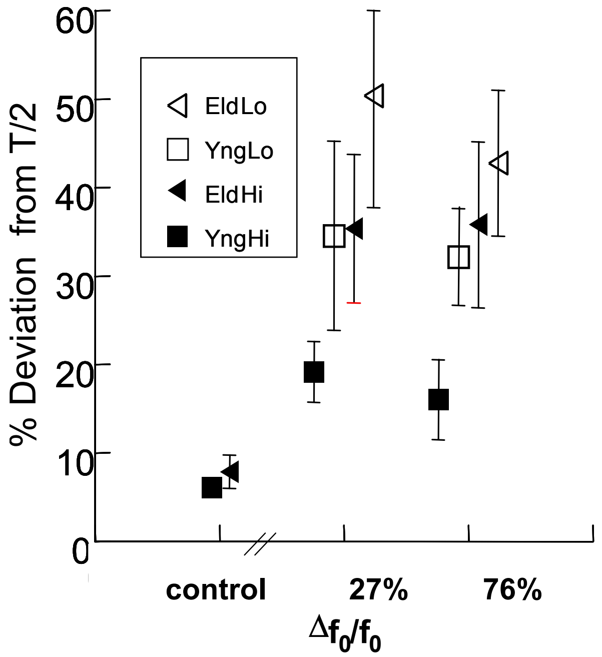

Figure 2. Results of Experiment 1: the effect of fundamental frequency f_0 difference. The rhythmic pattern had a total duration of 400 ms and the bursts in both the target and the distracter stream were modulated with a 100 percent modulation depth, i.e., exactly as shown in Figure 1. Average data and standard error of the mean of five young (squares) and seven elderly (triangles) subjects. The ordinate represents the percent deviation of the duration of the inter-burst time interval from rhythmic regularity at which both the interval between the first and the second burst and the interval between the second and the third burst would have a duration of T/2, i.e., 200 ms. The leftmost data points, labeled "control", stand for the condition of discrimination of the rhythmic pattern alone, without the presence of the distracter stream. The other two data points are data for the segregation condition in which the f_0 of the two streams was higher by 27 (~Major third) and 76 percent (~minor seventh) than the lower f_0 which was always 107 Hz. The filled symbols represent conditions in which the target pattern stream had the higher f_0 (i.e., either 136 or 189 Hz) and the distracter stream the lower f_0, whereas the unfilled symbols are for the conditions in which the f_0 of the target pattern stream had the lower f_0 and the distracter the higher f_0.

In Experiment 3, we examined the discriminability of the pattern rhythm at three different rates of modulation, the fastest corresponding to a subsyllabic and the slowest to a drawn-out syllabic rate. Data of our elderly and young listeners are shown in Figure 4 for the 100-percent modulation depth and the largest (76-percent) fundamental frequency difference, with the pattern appearing in the high-f_0 and the distracter in the low-f_0 stream. This experiment is analogous to examining the effect of speaking rate on the perception of prosody.

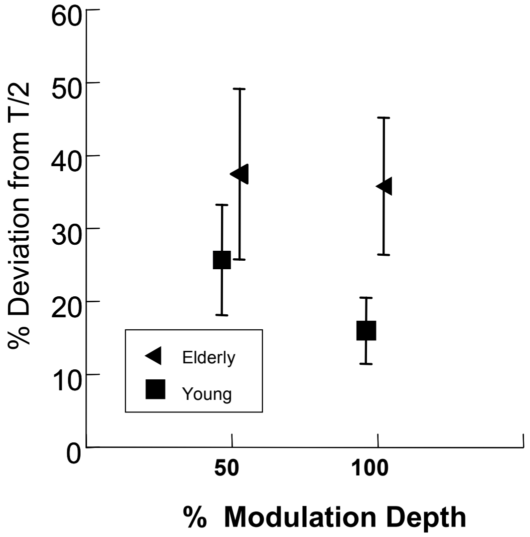

Figure 3. Results of Experiment 2: the effect of modulation depth. The target rhythmic pattern had a total duration of 400 ms, the fundamental frequency f_0 of the pattern was 189 Hz and that of the distracter 107 Hz, i.e., the f_0 separation of the two streams was 76 %. Average data and standard error of the mean of the young (squares) and the elderly (triangles) subjects. The ordinate is similar to that explained in Figure 2. The 100 percent modulation depth is the one illustrated in the diagram of Figure 1, whereas the 50 percent modulation depth refers to the condition in which the bursts, both in the pattern and the distracter streams, were riding on top of a continuous hum having a power 6 dB lower than the burst peaks.

4. DISCUSSION

From the results, several findings transpire. First, rhythmic discrimination of the pattern when it was presented alone, i.e., without the distracter in a different stream, was always easier than when the distracter was present. Therefore, the main finding is that the presence of the distracter always interfered with rhythmic discrimination in the target stream. In other words, the subjects were unable to segregate the temporal envelope structure of the two streams in any of the conditions investigated.

Furthermore, we also saw that, in the condition where the target stream had the higher fundamental frequency than the distracter was more than twice as accurate than when it was lower. A wider fundamental frequency separation between the streams (corresponding to pairing an average-pitched male voice with a rather low-pitched female voice) yielded somewhat better results than a narrower f_0 separation (corresponding to pairing an average-pitched male voice with a rather high-pitched male voice), but rhythmic perception even at the wider f_0 separation was poorer than in the control condition, i.e., when the distracter was absent. Reducing the modulation depth by one-half made rhythmic discrimination more difficult. The largest disruption of rhythmic discrimination was caused by increasing the modulation rate ("speaking rate") from 5 to 10 Hz; decreasing it to 2.5 Hz improved discrimination only for the elderly subjects.

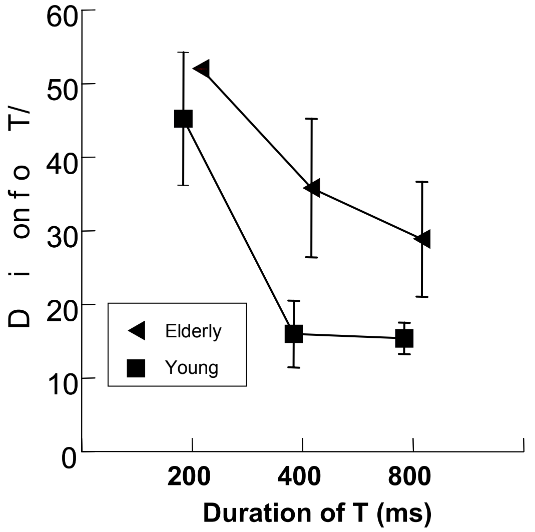

Figure 4. Results of Experiment 3: the effect of modulation rate. The fundamental frequency f_0 of the pattern was 189 Hz and that of the distracter 107 Hz, i.e., the f_0 separation of the two streams was 76 %; the modulation depth was 100 percent in both the target and the distracter streams. Average data and standard error of the mean of the young (squares) and the elderly (triangles) subjects. The ordinate is similar to that explained in Figure 2. The abscissa corresponds to the duration ***T*** of the total target pattern, i.e., to the separation between the first and the third bursts. Because the average duration between consecutive bursts was ***T***/2, the abscissa can also be viewed as depicting modulation rate from the fastest (10 Hz) to the slowest (2.5 Hz).

Unquestionably, the most consistent finding throughout all experiments was the large difference between the performance of the young and the elderly subjects. It has to be pointed out that none of the elderly subjects had a hearing loss worse than moderate and, given the 3-kHz upper bound of the frequencies used and the clearly suprathreshold levels at which the stimuli were presented, the elderly subjects' hearing loss could not have been a factor of their poor performance. In fact, when the pattern was presented alone, without the distracter, the elderly and the young subjects exhibited identical performances. Therefore, we must conclude that the ability to segregate two streams of harmonic complexes having different fundamental frequencies and different patterns of syllabic/subsyllabic rhythm deteriorates with age. Since the same elderly subjects also demonstrated a reduced ability to understand speech in babble noise or in reverberation, the possibility exists that the deficit of prosodic rhythm perception is related to, or possibly underlies, some of the age-related loss of the "cocktail-party" effect.

ACKNOWLEDGMENT

The authors wish to thank their colleague Dr. Brian Gygi for his assistance with the experiments and for contributing valuable ideas. The research was supported by a grant from the National Institute on Aging and by the Veterans Affairs Medical Research.

REFERENCES

[1] Greenberg, S., H.M.Carvey, L. Hitchcock, and S. Chang, "Beyond the phoneme A juncture-accent model for spoken language". *Proceedings of the Second International Conference on Human Language Technology Research*, 2002: p. 36-43.

[2] Blasko, D.G. and M.D. Hall, "Influence of prosodic boundaries on comprehension of spoken English sentences". *Percept Mot Skills*, 1998. **87**(1): p. 3-18.

[3] Raczaszek, J., B. Tuller, L.P. Shapiro, P. Case, and S. Kelso, "Categorization of ambiguous sentences as a function of a changing prosodic parameter: a dynamical approach". *J Psycholinguist Res*, 1999. **28**(4): p. 367-93.

[4] Price, P.J., M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation". *J Acoust Soc Am*, 1991. **90**(6): p. 2956-70.

[5] Avendano, C., H. Hermansky, M. Vis, and A. Bayya, "Adaptive speech enhancement based on frequency-specific signal-to-noise ratio estimates". *Proceedings of the IEEE IVTTA'96 Workshop*, 1996.

[6] Divenyi, P.L., "Dimensions of auditory segregation: What do they tell us about levels of auditory processing?", in *Physiological and Psychophysical Bases of Auditory Function*, D.J. Breebart, A.J.M. Houtsma, A. Kohlrausch, V.F. Prijs, and R. Schoonhoven, Editors. 2001, Shaker Publishing BV: Maastricht (the Netherlands). p. 468-476.

[7] Bregman, A.S., *Auditory scene analysis*. 1991, Cambridge, Mass.: Bradford Books (MIT Press).

[8] Divenyi, P.L., K.M. Haupt, and A.P. Algazi, "Auditory scene analysis in the elderly: Pairs of simultaneous complex tones segregated by temporal pattern or spatial location", in *Abstracts of the Twenty-first Midwinter Research Meeting, Association for Research in Otolaryngology*. 1998: St. Petersburg Beach, FL. p. 80.

[9] Levitt, H., "Transformed up-down methods in psychoacoustics". *Journal of the Acoustical Society of America*, 1971. **49**: p. 467-477.